图书情报工作
第65卷 第10期 2021年5月

网络社交平台中社群标签生成研究*

蒋武轩　易明　熊回香　童兆莉
华中师范大学信息管理学院　武汉 430079

摘　要：[目的/意义] 基于网络社交平台中社群话题及用户兴趣挖掘而生成的社群标签，能够提高社群定义的及时性与准确性，解决用户信息获取、网络社群选择的困难。[方法/过程] 通过对网络社群的深入分析，发现社群特征可根据社群话题及用户兴趣予以表征。首先，利用主题提取 BTM 模型对网络社群话题进行主题模型训练，从而得到网络社群话题预标签；其次，根据社群成员兴趣标签网络中不同类型的重要节点指标，利用 TOPSIS 多指标综合评价方法挖掘成员整体兴趣，从而得到网络社群成员兴趣预标签。综合两者结果生成社群标签并进行优化，且以"豆瓣小组"为例进行实证。[结果/结论] 基于社群话题及成员兴趣的社群标签生成模型能够准确地挖掘主要兴趣及近期关注点，社群整体的标签生成有利于网络用户兴趣群体的选择。

关键词：社群标签　标签生成　BTM　TOPSIS
分类号：TP393
DOI：10.13266/j.issn.0252-3116.2021.10.009

1 引言

随着网络技术及用户的不断增长，用户线上交流、分享、合作的本能不断显现[1]，各类网络社交平台逐步形成并快速发展，如豆瓣网、微博、微信等。网络用户根据个人需要或兴趣加入不同的网络社群与社群内其他用户交流、分享，网络社群已经是互联网用户的最大组织方式。但不同用户加入网络社群的目的不同，兴趣点也不同，在面对纷繁复杂的网络社群时，用户通常不知道某一社群所关注的重点，通常不知是否应该加入，是否符合自身需求与兴趣，只能根据片面的相关信息有选择性的加入大量相关网络社群，在经过一段时间的了解后再进行重点选择，这造成了用户的信息获取困难、效率低下等问题。同时，网络社群的关注点及兴趣也会随着外部环境的变化而改变。虽然搜索引擎可以根据网络社群的内容进行检索，但单一的内容不能全面的展现网络社群的整体特征，也不能发现网络社群的兴趣变化。因此，如何帮助用户准确了解网络社群的兴趣并及时展现网络社群关注点的变化，逐渐受到了学界和产业界的关注。

当前，国内外学者针对网络社群主要从概念、用户行为、信息传播与知识共享3个方面进行研究。在网络社群概念方面，H. Rheibgold 第一次提出网络社群的概念，认为网络社群是较多网络用户共同参与某一话题的讨论并形成一定的凝集关系[2]。其后陆续有学者进行研究，G. Siemens 认为具有共同兴趣的网络用户进行持续的互动和分享即为网络社群[3]；N. D. B. Navarro 又提出一个自发性社交网络（Spontaneous Social Network，SSN）的概念，通过对网络社群中的社群意识、归属感、社会有用性、成员忠诚度和社群的短暂性进行评估，发现该类网络社群拥有较好的社交感知虚拟环境[4]。在社群用户行为方面，T. Zhou 利用社会感知理论来确定影响用户持续性使用知识社群的因素，发现对结果期望的认知因素和系统、知识质量的环境因素显著影响用户的持续意图，进而影响持续使用行为[5]；刘佩分析研究了网络社群中的"小世界"网络关系，发现信息传播与点的入度和出度数有关[6]。邓卫华发现信息传播活动中接收、再传播和发布是最为常见的3种活动，个体社群用户的传播行为可分为接受型、扩散型和创造型3个层次，现以扩散型为主[7]。

* 本文系国家社会科学基金年度项目"融合知识图谱和深度学习的在线学术资源挖掘与推荐研究"（项目编号：19BTQ005）研究成果之一。
作者简介：蒋武轩（ORCID：0000-0001-9621-4318），博士研究生，E-mail：jiangchair@mails.ccnu.edu.cn；易明（ORCID：0000-0002-4864-6025），教授，博士生导师；熊回香（ORCID：0000-0001-9956-3396），教授，博士生导师；童兆莉（ORCID：0000-0003-1621-4356），博士研究生。
收稿日期：2020-12-09　修回日期：2021-03-18　本文起止页码：79-89　本文责任编辑：徐健

在社群信息传播和知识共享方面，C. C. Liao 对网络社群知识共享进行综合分析，认为使用动机、享乐动机、自我效能和共享文化能够激发用户对知识共享的态度[8]；C. Chen 以价值创造理论为切入点，认为知识共享有助于提高用户的共同创造价值，这些价值包括用户学习价值、社会综合价值和享乐价值，且后续会影响用户未来再参与的意愿[9]。同时，近几年有学者逐渐发现社群标签对网络社群的重要性，H. Xie 等提出利用多种关系提升社交平台的标签使用，结合资源的内容和标签对用户社群进行聚类，发现潜在社群[10]；李文根认为社区问答的内容主要以短文本为主，传统的文本处理方法对其并不适用，因此借助 Wikipedia（维基百科）作为外部知识库构建图模型的标签生成方法[11]；蒋武轩利用网络社群话题及成员兴趣标签构建社群标签动态生成模型，使用社群动态标签表征社群主要特征[12]。综上所述，国内外对于网络社群标签的研究较少，且主要是针对用户进行潜在社群推荐方面，而对社群整体标签研究方面还处于初始阶段。

本文在当前研究及前期网络社交平台中社群标签动态生成研究的基础上，以网络社群作为研究对象，将主题模型、复杂网络、管理决策相关方法技术相结合，构建网络社群标签生成模型。通过对网络社群的分析发现，网络社群标签可以从社群话题及社群用户近期兴趣两个方面来挖掘。因此，本文提出从社群话题及用户近期兴趣标签对社群标签进行动态生成。首先通过 BTM（Biterm Topic Model）模型提取网络社群的动态话题进行主题模型训练，从而得到网络社群话题预标签；其次根据网络社群活跃用户兴趣标签网络中不同类型的重要节点指标，利用 TOPSIS 多指标综合评价方法挖掘成员整体兴趣，从而得到网络社群成员兴趣预标签；最后在两者基础上，根据不同领域不同网络社交平台对最终动态标签进行合成。

2 社群标签生成模型

网络社群的特征可以从两个方面进行挖掘：一是社群的讨论话题，社群成员在社群中通过各种方式针对某些问题进行讨论与交流，代表着社群主要的关注点；二是社群活跃用户的近期兴趣，社群活跃用户是社群成员的主体，其近期兴趣能够较好地表征社群近期整体的相关兴趣。因此，本研究整合社群近期话题与活跃用户近期兴趣，构建社群标签生成模型，如图 1 所示：

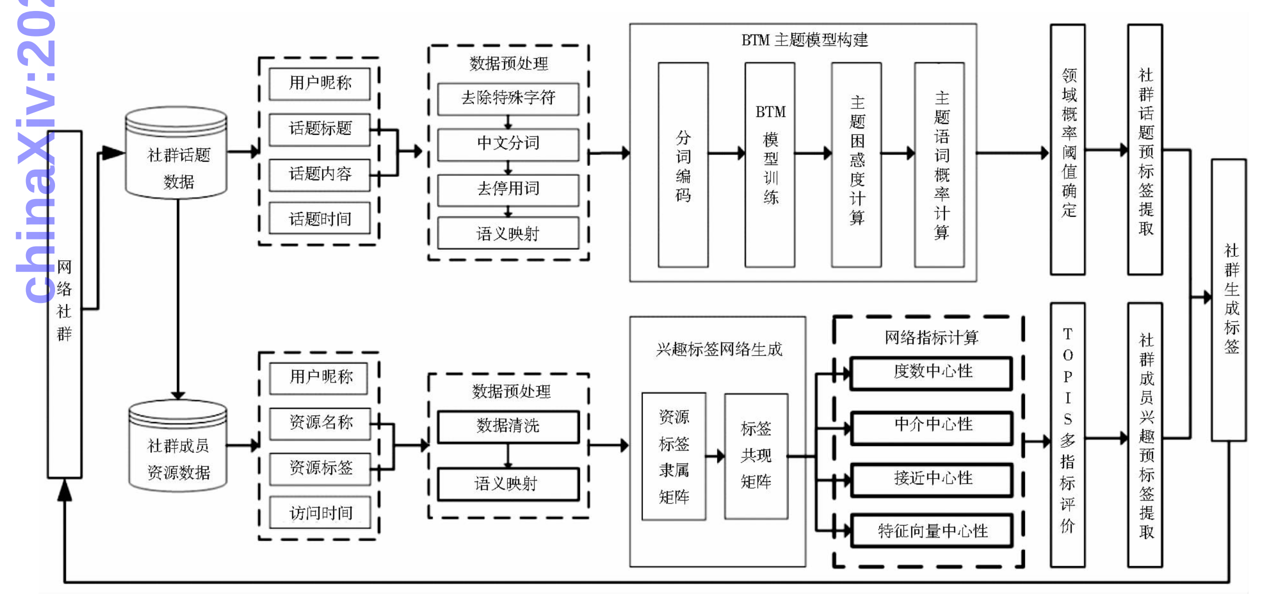


图 1 社群标签生成模型

该模型包括两个子模型，社群话题标签动态子模型与社群成员兴趣动态标签子模型。该模型将自动从社群中收集近期社群成员发表的讨论话题及参与话题成员近期观看资源及相关资源标签。首先，在依据子模型进行数据预处理的基础上，对社群话题数据经过 BTM 主题模型训练后提取相应阈值的主题语词作为话题预标签；其次，运用社会网络的思想对社群成员近期兴趣数据构建兴趣标签网络，采用社会网络中重要节点的 4 个经典度量指标对兴趣标签网络进行度量，再利用多指标评价方法 TOPSIS 将 4 个指标进行综合评价，提取成员兴趣预标签；最后，将两个子模型的预标签进行综合处理，确定最终社群标签。同时，该模型

蒋武轩，易明，熊回香，等. 网络社交平台中社群标签生成研究[J]. 图书情报工作,2021,65(10):79 - 89.

在一段周期内不断更新话题与成员兴趣信息,动态更改社群标签,最大程度表征社群特点及近期关注情况,及时、准确地表征社群特征,方便用户清楚地了解不同社群特点。

2.1 社群话题标签子模型

社群话题是用户根据自身需要发表的关于本社群主题的内容,通过对其进行主题提取能够表征社群的主题。针对社群话题分析数据较少的情况,本研究主要采用短文本主题模型 BTM 对近期话题标题及其内容进行主题提取。BTM(Biterm Topic Model)是由 X. Yan 教授等在 2013 年 5 月 IW3C2 会议上提出的专用于短文本的主题挖掘模型,该模型通过词共现的模式来加强主题模型的学习,并利用整个语料库的丰富信息抽样主题,以推断整个语料库全局的主题分布,能够有效解决文档级别的数据稀疏性问题。

针对获取的相关数据进行预处理操作,包括去除特殊字符、中文分词、去停用词、语义映射。在中文分词方面,本研究利用 Python 语言基于 ICTCLAS 中文分词系统,以每篇话题为一条记录对经过前述特殊字符去除处理的标签进行分词。该系统针对中文的分词结果准确性较高,并具有自定义词典的功能,能够根据分词需要添加新词,以提高分词的准确性,如某些资源的名称、涉及的人名等,若将其拆分将对后续提取主题造成干扰,因此将所有涉及到的资源名及人名等添加到自定义字典中,使其不再进一步拆分,以提高话题分词的准确性。而在语义映射方面,由于经过前期处理后许多语词存在语义相似的情况或文体不同的情况,通过计算语词间的语义相似度对其进行归一化处理,提高其后续分析的准确性。

在数据预处理的基础上,根据 BTM 主题模型对话题数据进行处理。首先,将话题数据作为社群文档集合 W,并将每个时间段的社群话题数据作为一个子文档集 w_i,i 表示不同子文档集,其中每一个文档都是一个话题。其次,对分词进行编码,并将每篇文档的分词结果用编码进行表示。此后,对社群文档集合 W 进行 BTM 模型训练,构建社群话题主题模型,并分别将每一子文档集作为新文档进行主题提取。其模型语词概率的计算方法为:

$$\mathrm{P}(\mathrm{B}|\alpha,\beta) = \prod_{i=1}^{N_B}\int\int\sum_{k=1}^{K}\theta_k\,\varphi_{k,w_{i,1}}\,\varphi_{k,w_{i,2}}\,d\theta d\varphi \quad 公式(1)$$

之后,通过主题模型困惑度确定主题数,困惑度计算如公式(2)所示[13-14]。

$$Perplexity(M) = \exp\left\{-\frac{\sum_K P(Z_K)\ \sum_{i=1}^{N_K} logP(W_i|Z_K)}{N}\right\} \quad 公式(2)$$

公式(2)中,K 表示主题数,$P(Z_k)$代表主题 K 的概率,N_K 表示主题 K 的主题词数,$P(W_i|Z_K)$表示主题 K 下第 i 个主题词的概率,N 代表该文档集中的所有语词数。同时,因为公式针对每个词都进行了计算,词频因素包含在其中,故公式中并未单独设定词频变量。在确定主题数 K 后,经过 1 000 次以上的迭代即可得到每一个子文档集下的主题 - 语词概率分布。最后,根据主题语词概率选取 TOP10 作为社群话题预标签。

2.2 社群活跃成员兴趣标签子模型

同一社群成员的主要兴趣是相似的,因此活跃社群成员的近期兴趣会存在相对较多的部分重合,通过不同用户兴趣间的关联即能够构建出社群成员的兴趣网络。该网络中将会存在一些能够表征多数用户共同兴趣的节点,这些节点具有较多的连接;也会存在一些只有较少用户感兴趣的节点,即节点的连接较少,一般而言这两种节点数量符合齐普夫(Zipf)定律。由于节点分布呈现出很大的异质性,并且节点的度也服从幂率分布,这样的网络符合社会网络的特征。该子模型将根据社群成员兴趣网络这一特点,利用资源标签表征用户兴趣,通过网络度量指标动态地挖掘社群成员的兴趣中心,即具有大量连接并主导网络运行的节点。

2.2.1 兴趣标签网络生成

(1)资源 - 标签隶属矩阵。根据社群活跃用户的资源标签,利用 R 语言编写算法构建"资源 - 标签隶属矩阵",该矩阵将所有资源与标签包含在同一矩阵中,有利于后续处理。矩阵中行表示具体某资源,列表示所有标签,而矩阵内的值表示该资源是否使用该标签。构建算法如下:

```
Source_Targ <- read.csv("资源标签数据.csv",header=F)
Targ <- read.csv("隶属矩阵框架.csv",header=F)
rnum = nrow(Source_Targ)
for (i in 1:rnum):
  source = as.character(Source_Targ[i,1])
  #获取资源在隶属矩阵中横坐标
  x = which(Targ[,1] == source)
  lnum = ncol(Source_Targ[i,1])
  for (j in 2:lnum):
    targ = as.character(Source_Targ[i,j])
    #获取资源对应标签在隶属矩阵中的纵坐标
    y = which(Targ[1,] == targ)
    if (sum(y)! =0){
      Targ[x,y] <- 1
```

```
        }else{
          next()
        }
    }
}
write.csv(Targ, file = "资源 - 标签隶属矩阵.csv", row.names = FALSE)
```

（2）标签共现矩阵。利用"资源 – 标签隶属矩阵"能够发现不同标签同时为同一资源进行标注，则这些标签间存在共现关系，通过"资源 – 标签隶属矩阵"，将其转换为标签共现矩阵，便于后续重要标签的挖掘。

2.2.2 资源标签网络度量指标

社会网络中衡量网络中节点重要性的指标有很多，笔者选取较为经典4个指标进行度量，分别为：

（1）度数中心性。度数中心性 $C_{D(i)}$ 是社会网络分析中刻画节点中心性最直接的度量指标，是指标签网络中某个节点的连接数，测量的是标签网络中标签与其他标签共现的数量，刻画了该标签节点与其他标签节点直接建立关联的能力，即度值。其能够体现某个标签控制其他标签的数量及能力。如果节点的 $C_{D(i)}$ 值越大即说明该标签与社群标签网络中其他标签关系十分密切，这个标签越重要。

（2）中介中心性。中介中心性 $C_{B(i)}$ 是指标签网络中经过某个标签并且连接其他标签的最短路径数量与这两个标签之间所有最短路径数量的比例，用以测量标签控制其他标签的信息交流能力。某标签的中介中心性值越大说明该标签在标签网络中对信息的协调能力越强，表示其处于标签网络中的枢纽位置。

（3）接近中心性。接近中心性 $C_{C(i)}$ 指一个标签与标签网络中其他标签的距离之和，能够体现标签网络中的某标签与其他标签之间的距离长短，探索网络中各个标签之间关系的强弱。其不仅要考虑标签节点的值，还要考虑标签节点在网络中所处的位置，更能反映标签网络的整体结构。

（4）特征向量中心性。特征向量中心性 $C_{E(i)}$ 更加关注标签节点间的相互影响即群体效应，标签节点可以通过与其它重要标签节点的连接间接地提高网络的地位。即一个标签节点是否重要，不仅与其自身有关，还与其连接的标签节点有关。

2.2.3 社群成员兴趣预标签

为了将不同评价指标进行综合，本研究引入多指标评价体系中经典的算法 TOPSIS，该算法是一种"逼近于理想值"的排序方法，适用于根据多项指标对多个方案进行比较选择的分析方法，在本研究中即根据2.2中的4个指标从社群成员兴趣标签网络计算出较优的标签。这种方法首先要确定各项指标的最优方案（即正理想值）与最坏方案（即负理想值），然后求出各个方案与正、负理想值之间的加权欧式距离，从而获得各个方案与最优方案的接近程度，作为评价方案的优劣标准。TOPSIS 算法步骤为：

步骤1：构建社群标签网络的决策矩阵 X；

步骤2：将决策矩阵 X 标准化，构建社群标签网络的标准化决策矩阵 Y；

步骤3：根据标准化决策矩阵 Y 确定正理想值 Y+ 和负理想值 Y–；

步骤4：计算各方案到正理想值 Y+ 的距离 D+ 和到负理想值 Y– 的距离 D–，进而计算各方

案的综合评价指数 Gi。

最后根据 TOPSIS 综合评价指数 Gi 排序选取 TOP10 作为社群成员兴趣预标签。

2.3 社群标签生成

在2.1与2.2中得到的子模型预标签基础上进行整合，生成社群标签。由于不同领域的社群情况不同，笔者认为不同领域的社群话题需要根据具体情况设置不同的概率阈值来进行预标签提取，以保证预标签的显著差异性。同时，剔除社群成员兴趣标签 TOP10 中对社群表征无意义的标签，最终产生成员兴趣预标签。

由于社群话题标签主要表征变化较少的整体兴趣，而社群成员兴趣标签表征变化较大的用户近期关注点，为了使社群标签能够更加准确地表征，笔者认为应根据不同领域设定两类标签整合的分配比例，由于一般社群整体特征标签变化较小且数量较少，而成员兴趣标签变化较大且数量较多，并且一般社群标签数量均为5个左右，因此为了兼顾两类标签的因素，将比例设定为2:3能够适合大多数领域社群。若话题预标签数量较少，则由成员兴趣标签进行补充；若话题预标签与成员兴趣预标签存在重叠情况，则将该标签设定为 Top1 成员兴趣，其他标签选取顺序依次顺延。据此，生成最终社群标签。

3 实证研究

随着网络社交平台的发展，网络社群的爆炸式增长，其中最为经典的网络社群即"豆瓣网 – 小组"，豆瓣用户可自由创建小组，小组涉及内容包罗万象，如电影、读书、音乐、手工艺、陶艺、文具等涉及生活的方方面面，随着小组逐渐增多，目前几乎囊括所有类别的小组，但由于是自由创建，相同类别的小组被大量重复创

蒋武轩，易明，熊回香，等.网络社交平台中社群标签生成研究[J].图书情报工作,2021,65(10):79－89.

建,使得同一类别的小组更加繁杂。同时,“豆瓣网－小组”既提供了小组发帖讨论的功能,又基于豆瓣网大量的资源允许用户进行资源的浏览与标签的设定。“豆瓣网－小组”作为较为流行的网络社群,网络用户较多,数据较为丰富,但又存在较为明显的问题,因此本研究以“豆瓣网－小组”为实证研究对象,从中获取数据并对提出的模型加以验证。

3.1 数据收集与整理

本研究采用 Python 编写网络爬虫,以一个月为间隔,分别爬取同类别豆瓣网－小组相同时间段及不同类别豆瓣网－小组不同时间段的数据进行对比验证,同时为了便于实证结果的验证,本文选择性地爬取小组类型能够被其名称明显标识的小组数据。采集数据包括:某一时间段特定小组所有帖子标题及内容、帖子中涉及到的豆瓣用户昵称、涉及的用户在时间段内浏览的资源名称及资源的标签。具体数据如下:

(1)2018 年 1 月、2018 年 2 月“佳片推荐”小组数据,同时为了验证实证结果也爬取了该小组较长时间间隔即 2018 年 12 月的数据;

(2)相同类型的小组“一个人看电影”2018 年 1 月、2018 年 2 月、2018 年 12 月时间段的数据;

(3)不同类型的小组“买书 读书 一起来吧”2018 年 12 月、2019 年 1 月时间段的数据。

其中,小组话题部分数据如表 1 所示,成员兴趣资源及标签部分数据如表 2 所示:

表 1　部分小组话题数据

用户昵称	话题标题	话题内容
000000	一日情人 L' amant d' un jour	https://www.douban.com/doubanapp/dispatch? uri=/review/9055612/&dt_dapp=1
37	『求电影名字』	一个女的,带着小孩去雪山一个地方租房子住。她老公不在她身边,什么事情只能给她老公打电话。房东帮过她很多,他们慢慢喜欢上对方。后来,女的去找她老公,走了。几年后,女主角又回到这个地方找他。男主角以前为了她摔断了腿。因为机缘巧合,这次女的没有能见到男主角。男主角知道后,开车不顾大雪封山,下山去找她,终于相遇,两个人如愿以偿的在一起了
123	有看过万能钥匙的吗	恐怖吗、本人胆小又想看
5492697	看过发条橙的吗	来探讨一下人性呀
---	看过最多遍的电影	说一部你看过最多遍的电影
!	最近剧荒	求推荐动作,悬疑,犯罪,都可以! 不要脑残片,谢谢
。	有关战争的电影	拯救大兵瑞恩,血战钢锯岭,狂怒。有关战争类的求推荐。
Ace'	求片名	之前看了一部电影,讲得是一个日本人搬到了一座美国公寓里,然后门房是一个美国老寡妇,门房的朋友是一个小孩,后来门房差点和日本人谈恋爱,但是因为一些误会不愉快了还是怎么了,最后寡妇被车撞死了,这是啥电影
AN	大家认为电影开罗时间怎么样? 多图预警	废话不多说,上图 原来还有平淡如水的艳遇,这部电影告诉我的
Arphy	求感人电影	求一部感人到哭的电影,谢谢谢谢
blood 雪之族	谁介绍下赛博朋克,反乌托邦题材电影?	就是像《银翼杀手》这样的电影…
blood 雪之族	有没有什么好的儿童电影?	比如《伴我同行》这样的
blood 雪之族	刚刚看完《分裂》,谁能给我稍微解析一下吗?	有点没看懂,还有男主角最后是分裂出新的人格了吗? 怎么那么厉害? 真的像野兽那样爬来爬去,也太超现实了吧。最后他不杀女主又是啥意思? 这部电影是不是涉及到一些心理学知识?
Bon Homme	求推荐好看的电影或剧	推荐时请注明 名字 哪里可以下载 什么类型 最近剧荒了 笔芯 靴靴
Bourne	豆瓣影人里面的照片咋没了	你们还能看吗?
chuchu	一起来看电影啊	每周三分享一部电影,大家可以在群里分享观影感受,群里会不定期发福利红包,也可以互相分享推荐喜欢的电影电视剧,讨论近期追的剧和电影~
cocojamboo	说一个让你震撼的电影片段	自己先开个头吧,大家保持队形《漂流欲室》－ 金基德导演的, 女主把鱼钩放入下体,然后拽鱼线。被惊到了。。。豆瓣链接: https://movie.douban.com/subject/1305088/
…	…	…
Justseven	无问西东。	谈谈感受吧

由于实证数据是通过爬虫自动抓取,数据类型多样化,因此存在以下问题:①存在不同外文资源但中文名称相同或同名资源的现象,整理过程中通过在资源名称后加注年份进行区分;②存在社群成员参与了最近的话题讨论但未有资源的现象,即缺少该成员的资源标签数据,针对这部分数据在整理中保存话题但在资源数据中进行剔除。经过对数据进行补充和梳理,共有 2 113 篇话题讨论,涉及 1 578 名成员,共计 4 696 个资源,8 214 个资源标签,见表 3。

表 2　部分社群活跃用户浏览资源标签数据

资源	标签
画廊外的天赋	纪录片 艺术 传记 电影 文艺
一日情人	爱情 文艺 黑白 女性 戛纳
年轻气盛	文艺 人生 人性
偷香	贝托鲁奇 爱情 青春 情色 意大利电影 Bernardo-Bertolucci
水牛城 66	爱情 独立电影 黑色幽默 黑色 独立
蓝白红三部曲之白	文艺 爱情 波兰 经典 人性
另一个波琳家的女孩	历史 宫廷 传记 爱情 女性
血战钢锯岭	战争 信仰 真实事件改编 二战 人性 历史 军事
巴黎淘气帮	喜剧 儿童 童年 成长 温情 搞笑 家庭
妖猫传	唐朝 奇幻 古装 魔幻 悬疑 猫
白夜行	東野圭吾 日本电影 堀北真希 悬疑 东野圭吾 白夜行
妖铃铃	喜剧 烂片 搞笑 香港 开心麻花 国产
推销员之死	达斯汀・霍夫曼 DustinHoffman 推销员之死 美国电影 施隆多夫
巴尼的人生	加拿大 爱情 人生 传记 温情
…	…
樱桃小丸子:来自意大利的少年	小丸子 动画 童年 剧场版 温情 日本动漫 经典

表 3　实证数据统计

社群名称	时间	话题数	用户数	资源数	标签数
佳片推荐	2018 年 01 月	323	264	759	1 204
	2018 年 02 月	238	210	846	1 298
	2018 年 12 月	427	320	870	1 214
一个人看电影	2018 年 01 月	268	209	657	1 168
	2018 年 02 月	160	136	380	594
	2018 年 12 月	389	313	874	1 340
买书 读书 一起来吧	2018 年 12 月	159	66	140	729
	2019 年 01 月	149	60	170	667
总计		2 113	1 578	4 696	8 214

同时,经过对话题发布时间进行统计,发现话题发布时间主要集中于数据收集前 5 天,其他时间的话题主要是以前的话题有了新的回复,可以得出社群活跃度较高,每天都有成员进行话题讨论,社群内容更新较为快速,具有较好的研究价值。

3.2　社群话题标签生成

对话题数据进行预处理操作,分别对话题数据进行去除特殊字符、中文分词、去停用词、语义映射后,得到社群话题预处理结果。在此基础上,将话题数据作为社群文档集合 W,并将每个时间段的社群话题数据作为一个子文档集,将其分为 8 个子文档集合(“佳片推荐”2018 年 1 月话题数据为 w1,2018 年 2 月话题数据为 w2,2018 年 12 月话题数据为 w3;“一个人看电影”2018 年 1 月话题数据为 w4,2018 年 2 月话题数据为 w5,2018 年 12 月话题数据为 w6;“买书 读书 一起来吧”2018 年 12 月话题数据为 w7,2019 年 1 月话题数据为 w8),文档集中的每一个文档 Di 都是一个话题。对文档集进行预处理,将分词进行编码,并将每篇文档的分词结果用编码进行表示,如表 4 所示:

表 4　话题分词部分编码表示

文档	预处理结果	分词编码表示
D1	情人	0
D2	电影 名字 小孩 雪山 租房子 老公 老公 房东 喜欢 对方 老公 女主角 男主角 摔断了腿 机缘巧合 男主角 男主角 开车 不顾 大雪 封山 下山 相遇 如愿以偿	1 2 3 4 5 6 6 7 8 9 6 10 11 12 13 11 11 14 15 16 17 18 19 20
D3	万能钥匙 恐怖 胆小	21 22 23
D4	探讨 人性	24 25
D5	电影 电影	1 1
…	…	…

利用 Python 调用编写程序对社群文档集合 W 进行 BTM 模型训练,构建出社群话题主题模型。根据公式(1)计算文档集 W1 不同主题数 K 的困惑度得到困惑度曲线,可以看出不同主题模型的 K 值越大,困惑度越低,如表 5 所示,但困惑度只在 0.001 级别内进行波动,并无较为显著的差异,因此设置主题数 K = 1,根据 BTM 模型在经过 1 000 的迭代之后得到每一子文档集下的主题 – 语词概率分布,如表 6 所示。

表 5　主题困惑度值

主题数	$-\{\sum_K P(Z_K)\ \sum_{i=1}^{N_K} logP(W_i \mid Z_K)\}/N$	主题困惑度
1	0.018 396 830	1.018 567 094
2	0.017 716 041	1.017 873 901
3	0.017 102 683	1.017 249 771
4	0.016 967 775	1.017 112 545
5	0.016 421 836	1.016 557 415
…	…	…

表 6 展示了“佳片推荐”3 个时间段(w1 – 3)、“一个人看电影”3 个时间段(w4 – 6)及“买书 读书 一起来吧”2 个时间段(w7 – 8)子文档集的“主题 – 语词”概率分布,其中每一行表示一个子文档集的主题 – 语词及其概率,如第 1 行文档集 w1 中,该主题下共有 10 个语词,其中“电影”这一语词表征该文档集的主题概率为 0.043 723。通过对各子文档集的主题 – 语词概率进行比较,发现“佳片推荐”社群中“电影”“推荐”等语词的概率与其他语词相比具有较为明显的差别,而文档集 w7“书籍”“买书”语词概率也具有显著区别,且在同一社群不同的子文档集中概率都比较高,较为稳定,表征了社群的主要兴趣。

蒋武轩，易明，熊回香，等. 网络社交平台中社群标签生成研究[J]. 图书情报工作,2021,65(10):79－89.

表 6　子文档集主题－语词概率分布

文档集 W_1	语词	电影	推荐	蝶衣	文工团	霸王	感动	世界	喜欢	一辈子	政治家
	概率	0.043 723	0.013 092	0.012 152	0.009 982	0.008 535	0.007 667	0.006 763	0.006 437	0.006 267	0.006 004
文档集 W_2	语词	电影	评分	男主	人数	女主	父亲	妻子	水形物语	蒂姆	儿子
	概率	0.018 638	0.011 022	0.010 291	0.010 237	0.008 989	0.008 989	0.007 491	0.006 242	0.005 993	0.005 493
文档集 W_3	语词	电影	推荐	喜欢	名字	妻子	工作	美好	情节	资源	家庭
	概率	0.107 958	0.043 027	0.013 422	0.012 435	0.010 856	0.010 264	0.007 855	0.008 488	0.008 093	0.007 698
验证数据文档集 W_4	语词	孩子	工作	电影	父母	水形物语	方式	老师	方法	爸爸	老板
	概率	0.019 891	0.006 816	0.006 445	0.006 057	0.005 922	0.005 686	0.004 724	0.004 724	0.004 707	0.004 252
文档集 W_5	语词	电影	生活	王彩玲	热爱	柏舟	感动	影片	进群	喜欢	分享
	概率	0.023 764	0.012 067	0.011 156	0.008 766	0.006 517	0.005 578	0.005 464	0.005 408	0.005 208	0.004 668
文档集 W_6	语词	电影	瑜伽	喜欢	老师	挽回	视频	分手	学习	济公	故事
	概率	0.025 442	0.021 329	0.008 265	0.008 226	0.005 308	0.004 936	0.004 642	0.004 524	0.004 407	0.004 309
文档集 W_7	语词	书籍	买书	书店	阅读	优惠券	京东	宇宙	封面	外星	印刷
	概率	0.083 162	0.030 079	0.012 621	0.012 007	0.011 461	0.010 916	0.010 234	0.009 551	0.009 347	0.009 142
文档集 W_8	语词	买书	书籍	京东	活动	自营	优惠券	参加	世界	名著	中国
	概率	0.036 611	0.031 007	0.025 381	0.023 510	0.020 566	0.019 501	0.018 948	0.013 345	0.010 556	0.010 247

3.3　社群成员兴趣标签生成

3.3.1　用户兴趣标签网络构建

在对爬取的不同资源出现的频率及标签词频进行整理与统计的基础上，利用算法 1 构建“资源－标签隶属矩阵”，部分矩阵数据如表 7 所示。资源－标签隶属矩阵展现了用户标注某一资源的常用标签，矩阵中行为资源名称，列为用户标签，矩阵中数值 1 表示该列标签是该行资源的用户常用标签，数值 0 则表示用户并未使用该标签标注对应资源。通过构建“资源－标签隶属矩阵”将资源与其标签数据进行整合，作为下一步构建“标签共现矩阵”即标签网络的基础。

3.3.2　兴趣标签网络

基于表 7，根据同一资源中标签间的共现关系，利用 Matlab 编写程序生成资源标签的共现矩阵，以“佳片推荐”1 月数据共 1 204 个标签为例，部分结果如表 8 所示。表中行列均是资源标签，数值 1 表示行列标签存在共现关系，数值 0 则表示不存在，而标签共现矩阵则为兴趣标签网络。

表 7　资源－标签隶属矩阵

标签 资源	纪录片	爱情	文艺	贝托鲁奇	历史	战争	喜剧	唐朝	東野圭吾	达斯汀·霍夫曼	…	Jude-Law
画廊外的天赋	1	0	1	0	0	0	0	0	0	0	…	0
一日情人	0	1	1	0	0	0	0	0	0	0	…	0
年轻气盛	0	0	1	0	0	0	0	0	0	0	…	0
偷香	0	1	0	1	0	0	0	0	0	0	…	0
水牛城 66	0	1	0	0	0	0	0	0	0	0	…	0
…	…	…	…	…	…	…	…	…	…	…	…	…

表 8　标签共现矩阵

标签 资源	纪录片	爱情	文艺	贝托鲁奇	历史	战争	喜剧	唐朝	東野圭吾	达斯汀·霍夫曼	…	Jude-Law
纪录片	0	0	1	0	1	0	1	0	0	0	…	0
爱情	0	0	1	1	1	1	1	0	0	0	…	1
文艺	1	1	0	0	1	1	1	0	0	0	…	1
贝托鲁奇	0	1	0	0	0	0	0	0	0	0	…	0
历史	1	1	1	0	0	1	0	0	0	0	…	0
…	…	…	…	…	…	…	…	…	…	…	…	…

3.3.3　标签网络重要节点指标计算

按照模型提出的 4 种兴趣标签网络重要节点度量指标：度数中心性、中介中心性、接近中心性、特征向量中心性。根据其各自的计算方法进行计算，利用 UCI-NET 计算兴趣标签网络各个标签节点的度量指标，为后续综合评价奠定基础。以“佳片推荐”社群 2018 年 1 月为例，部分数据如表 9 所示：

表 9　社群兴趣标签网络节点度量指标部分数据

标签 ID	标签名	度数中心性	中介中心性	接近中心性	特征向量中心性
1	纪录片	145	45 880	7 199	0.096
2	爱情	348	116 942.4	6 886	0.23
3	文艺	157	26 555.73	7 080	0.159
4	贝托鲁奇	5	0	7 966	0.01
5	历史	70	11 818.42	7 348	0.073
…	…	…	…	…	…

3.3.4　TOPSIS 兴趣标签网络综合评价

基于 TOPSIS 的多指标决策网络节点的重要性综合评价方法将标签网络中的每个标签节点看作一个方案,将多种评价指标看作各标签节点方案的属性,借此将标签节点重要性问题转换为多属性方案决策问题。本节利用 TOPSIS 综合评价方法探究兴趣标签网络中的重要标签节点。

(1)标准化决策矩阵构建。根据 3.3.3 中计算出的 4 种兴趣标签网络重要度指标,构建决策矩阵 X。其中行表示每个标签节点,列为每个节点的 4 种重要度指标,以“佳片推荐”社群 2018 年 1 月为例,决策矩阵 X 内容与表 9 内容相同。

同时由于不同指标的量级不同,为了方便比较因此需要对决策矩阵进行归一化处理,得到标准化决策矩阵 Y,如表 10 所示:

表 10　标准化决策矩阵 Y 部分结果

标签＼指标	度数中心性	中介中心性	接近中心性	特征向量中心性
纪录片	0.416 666 667	0.392 329 833	0.004 982 686	0.417 391 304
爱情	1	1	0.004 766 048	1
文艺	0.451 149 425	0.227 083 745	0.004 900 322	0.691 304 348
贝托鲁奇	0.014 367 816	0	0.005 513 554	0.043 478 261
历史	0.201 149 425	0.101 061 847	0.005 085 814	0.317 391 304
…	…	…	…	…

(2)正负理想值确定。根据标准化决策矩阵 Y 确定正理想值 Y+和负理想值 Y−。鉴于研究中基于网络结构的各中心性指标来考察社群标签节点的重要性,各指标之间并无权重高低之分,因此研究中并未对各属性设定权重向量,以“佳片推荐”社群 2018 年 1 月为例,正、负理想值如下所示:

$Y^{+}=\{1, 1, 1, 1\}$

$Y^{-}=\{0.008\ 620\ 689\ 655\ 172\ 41, 0, 0.004\ 766\ 047\ 689\ 546\ 6, 0\}$

(2)标签节点正负理想值距离计算。计算各标签节点到正理想值 Y+的距离 D+和到负理想值 Y−的距离 D−,部分结果如表 11 所示:

表 11　标签节点的正负理想值距离部分结果

标签＼距离	D^{+}	D^{-}
纪录片	1.727 087 993	0.995 233 952
爱情	1.992 578 894	0.001 484 631
文艺	1.727 087 993	0.995 233 952
贝托鲁奇	1.979 901 443	0.026 102 999
历史	1.982 051 044	0.021 757 772
…	…	…

(4)各标签节点贴近度计算。根据标签节点的正负理想值,计算各标签节点与理想方案的贴近度 G,并根据贴近度 G 降序排序,以“佳片推荐”社群 2018 年 1 月为例,部分结果如表 12 所示,整体过程数据如图 2 所示。

表 12　“佳片推荐”2018 年 1 月标签节点贴近度 G(TOP10)

标签名	爱情	喜剧	人性	经典	悬疑	美国电影	犯罪	文艺	香港	…
贴近度 G	0.634 417	0.601 772	0.450 864	0.437 728	0.425 342	0.407 363	0.391 071	0.376 793	0.370 632	…

同时,对 3 个社群其他日期的成员兴趣标签数据进行处理,“佳片推荐”标签部分结果如表 13 所示,“一个人看电影”如表 14 所示,“买书 读书 一起来吧”见表 15。

3.4　社群标签生成

将 3.2 节生成的社群话题标签与 3.3 节生成的社群活跃成员兴趣动态标签进行整合,生成社群动态标签。

通过对 3.2 节生成的社群话题标签进行分析,笔者认为在电影领域社群话题中概率大于 0.01 的语词与其他语词相比具有显著性差异,且较为稳定。因此本研究选取生成社群话题标签概率大于 0.01 标签作为话题预选标签,如“佳片推荐”社群 2018 年 1 月 25 日话题预选标签为“电影”“推荐”“蝶衣”;而书籍领域社群话题语词概率大于 0.03 则具有一定显著性。同时,对于生成的社群活跃成员兴趣标签,根据社群成员兴趣标签的贴近度 G 值排序(见表 12、13、14、15),由于国别、年份、资源类型(如中国、剧情、电视剧、2017 年)等标签对社群表征意义不大,因此对 TOP10 标签予以剔除后作为成员兴趣预选标签。

蒋武轩，易明，熊回香，等.网络社交平台中社群标签生成研究[J].图书情报工作,2021,65(10):79－89.

	标准化决策矩阵Y					
1						
2						
3	序号	标签	度数中心性	接近中心性	中介中心性	特征向量中心性
4	1	纪录片	0.416666667	0.004982686	0.392329833	0.417391304
5	2	爱情	1	0.004766048	1	1
6	3	文艺	0.451149425	0.004900322	0.227083745	0.691304348
7	4	贝托鲁奇	0.014367816	0.005513554	0	0.043478261
8	5	历史	0.201149425	0.005085814	0.101061847	0.317391304
9	6	战争	0.298850575	0.005019369	0.13195392	0.430434783
10	7	喜剧	0.890804598	0.004788196	0.824406915	0.952173913
11	8	唐朝	0.017241379	0.005590382	0	0.047826087
12	9	東野圭吾	0.014367816	0.005649213	0	0.02173913
13	10	达斯汀·霍夫	0.011494253	0.005740575	0	0.013043478
14	11	加拿大	0.106321839	0.005130803	0.016850103	0.260869565
15	12	人性	0.57183908	0.004862947	0.354389292	0.808695652
16	13	科幻	0.387931034	0.004964691	0.216404311	0.547826087
17	14	女性	0.298850575	0.004974381	0.108927435	0.543478261
18	15	短片	0.24137931	0.005054668	0.112503691	0.347826087
19	16	英国电影	0.146551724	0.005112116	0.050522851	0.239130435
20	17	国产电视剧	0.103448276	0.005186174	0.015358489	0.2
21	18	犯罪	0.485632184	0.004887171	0.293030041	0.673913043
22	19	传记	0.336206897	0.004959154	0.174812444	0.47826087
23	20	人生	0.278735632	0.004982686	0.080128479	0.517391304
1201	…	…	…	…	…	…
1202	1198	秋天的童话	0.020114943	0.005465105	0	0.060869565
1203	1199	桂纶美	0.020114943	0.005530166	0	0.043478261
1204	1200	半生缘	0.020114943	0.005529474	0	0.043478261
1205	1201	沈殿霞	0.020114943	0.005553698	0	0.043478261
1206	1202	赵薇	0.020114943	0.005504557	0	0.065217391
1207	1203	同性恋	0.020114943	0.005518399	0	0.052173913
1208	1204	Jude-Law	0.017241379	0.005491406	0	0.052173913

标签节点的正负理想值距离						标签节点贴近度
		Y+={1,1,1,1}		Y-={0.0086206896551 7241,0.0047660476895466,0,0}		G
序号	标签	(D+)^2	D+	(D-)^2	D-	
1	纪录片	2.039033157	1.427947183	0.494639765	0.703306309	0.329996555
2	爱情	0.99049062	0.995233952	2.982832937	1.727087993	0.634417247
3	文艺	1.984152865	1.408599611	0.725300429	0.851645718	0.376793486
4	贝托鲁奇	3.87540793	1.96860558	0.001923947	0.043862825	0.021795535
5	历史	2.902060911	1.703543633	0.148018153	0.384731274	0.184234017
6	战争	2.559505506	1.599845463	0.286919389	0.535648569	0.250831218
7	喜剧	1.035490437	1.017590505	2.364530371	1.537702953	0.601771569
8	唐朝	3.861300157	1.965019124	0.00236233	0.048603811	0.024137494
9	東野圭吾	3.917198618	1.979191405	0.000506399	0.022503316	0.011242132
10	达斯汀·霍夫曼	3.939778592	1.984887551	0.000179339	0.013391766	0.006701649
11	加拿大	3.301322893	1.81695429	0.077882504	0.279074369	0.133144348
12	人性	1.627030067	1.275550888	1.096795393	1.047279998	0.450863644
13	科幻	2.183207137	1.47756798	0.490820625	0.700585915	0.321642064
14	女性	2.484108913	1.576105616	0.391467236	0.625673426	0.28416721 8
15	短片	2.778402075	1.666853945	0.187816726	0.433378271	0.206347788
16	英国电影	3.198605213	1.788464485	0.078761013	0.280643926	0.135635197
17	国产电视剧	3.402978448	1.84471636	0.049228331	0.221874583	0.107362603
18	犯罪	1.860963019	1.364171184	0.767565375	0.876108084	0.391070924
19	传记	2.383873793	1.543979855	0.36660561	0.605479653	0.281689258
20	人生	2.589356512	1.60914776	0.347076464	0.589131958	0.267996813
…	…	…	…	…	…	…
1198	秋天的童话	3.831240357	1.95735545	0.00383771	0.061949257	0.030678509
1199	桂纶美	3.864078815	1.965726027	0.002023061	0.044978449	0.02236949 8
1200	半生缘	3.864080191	1.965726378	0.00202306	0.044978437	0.02236948 8
1201	沈殿霞	3.86403201	1.965714122	0.002023097	0.044978855	0.022369828
1202	赵薇	3.823014438	1.955253037	0.004385971	0.066226667	0.03276148
1203	同性恋	3.847542671	1.961515402	0.002854801	0.053430339	0.026517011
1204	Jude-Law	3.853236141	1.962966159	0.00279696	0.05288629	0.026235199

图 2　TOPSIS 兴趣标签网络综合评价“佳片推荐”2018 年 1 月整体过程数据

表 13　“佳片推荐”不同时间段兴趣标签网络标签节点贴近度 G

2018 年 2 月	贴近度 G	2018 年 12 月	贴近度 G
爱情	0.769 116 102	青春	0.635 646 042
喜剧	0.649 910 975	爱情	0.593 271 224
人性	0.461 290 09	喜剧	0.423 270 475
经典	0.437 751 639	人性	0.421 112 733
文艺	0.381 158 667	经典	0.409 195 616
电影	0.369 108 66	搞笑	0.373 283 624
温情	0.368 838 235	文艺	0.373 054 995
青春	0.362 222 011	荣光荣	0.364 353 958
美国电影	0.359 661 594	李维	0.364 353 958
动画	0.355 731 52	土耳其	0.364 338 716
…	…	…	…

表 14　“一个人看电影”不同时间段兴趣标签网络标签节点贴近度 G

2018 年 1 月	贴近度 G	2018 年 2 月	贴近度 G	2018 年 12 月	贴近度 G
爱情	0.503 82	剧情	0.368 35	美国	0.449 83
美国	0.446 56	美国	0.337 77	剧情	0.382 94
喜剧	0.320 61	爱情	0.335 11	喜剧	0.332 78
人性	0.312 98	2017 年	0.25	爱情	0.321 07
文艺	0.301 53	人性	0.248 67	2018 年	0.307 69
经典	0.274 81	动画	0.227 39	英国	0.260 87
剧情	0.225 19	文艺	0.219 41	青春	0.257 53
动作	0.221 37	悬疑	0.214 1	人性	0.209 03
英国	0.206 11	喜剧	0.188 83	动作	0.202 34
2017 年	0.206 11			科幻	0.200 67
…	…	…	…	…	…

表 15　“买书 读书 一起来吧”不同时间段兴趣标签网络标签节点贴近度 G

2018 年 12 月	贴近度 G	2019 年 1 月	贴近度 G
历史	0.031	文学	0.032
文学	0.019	小说	0.027
中国	0.018	外国文学	0.026
外国文学	0.018	历史	0.015
小说	0.014	随笔	0.012
随笔	0.011	中国文学	0.011
近代史	0.01	国学	0.01
读库	0.01	写作	0.009
文化	0.009	古典文学	0.009
英国	0.009	日本	0.009
…	…	…	…

因此，根据两类标签整合的分配比例，由于“豆瓣网－小组”标签规定为 5 个，则本研究中社群话题标签选取 2 个，社群成员兴趣标签选取 3 个，根据表 6 与表 12 以及上述的分析阐述，豆瓣电影兴趣小组“佳片推荐”（2018 年 1 月）的社群话题标签概率超过 0.01 的前两个为“电影”“推荐”；社群活跃成员兴趣标签贴近度 G 值最高前 3 个为“爱情”“喜剧”“人性”，因此豆瓣电影兴趣小组“佳片推荐”在该时间段的动态标签生成结果如图 3 所示：

“佳片推荐”生成标签	电影	推荐	爱情	喜剧	人性

图 3　“佳片推荐”小组标签生成结果

小组标签	电影	电视	导演	编剧	演员

图 4　“佳片推荐”小组原标签

对比分析图 3 的模型生成的小组标签结果与图 4 所示的该小组原标签，发现动态生成的标签既能够较为准确地反映出社群的特征，同时对社群短期的兴趣也有较好地揭示，将会方便用户的社群选择。

4 实证研究结果分析

本研究共抓取“佳片推荐”3 个时间段，“一个人看电影”3 个时间段，“买书 读书 一起来吧”2 个时间段，同一社群不同时间点、同类型社群相同时间点及不同类型社群的豆瓣兴趣小组话题及活跃成员兴趣标签数据。本节通过对这 3 种情况进行比较分析，并对模型效果进行验证。各社群动态标签生成结果如表 16 所示：

表 16 社群标签动态生成结果

社群名称	原标签	标签时间段	社群动态标签				
佳片推荐	电影 电视 导演 编剧 演员	2018.01	电影	推荐	爱情	喜剧	人性
		2018.02	电影	评分	爱情	喜剧	人性
		2018.12	电影	推荐	青春	爱情	喜剧
一个人看电影	电影 一个人生活 单身	2018.01	孩子	爱情	喜剧	人性	文艺
		2018.02	电影	生活	爱情	人性	动画
		2018.12	电影	瑜伽	喜剧	爱情	青春
买书 读书 一起来吧	买书 读书 聊天 书友 书讯	2018.12	书籍	买书	历史	文学	小说
		2019.01	买书	书籍	文学	小说	历史

实证研究的前提是用户对 3 个社群情况一无所知，且其社群名称不能标识其社群类型。在此情况下，根据表 16 可以发现，本研究生成的社群动态标签能够清晰地挖掘其社群特征。并且，不同类型社群间主要特征并不相同，“佳片推荐”与“一个人看电影”主要关注的是“电影”，而“买书 读书 一起来吧”关注的是“书籍”与“买书”。

同时，“佳片推荐”与“一个人看电影”在社群话题主要特征方面仍有很大不同，“佳片推荐”主要是对电影进行“推荐”“评分”，而“一个人看电影”则更具生活气息，生成标签主要为生活元素如“爱情”“生活”“瑜伽”。由此可以认为，模型在表征社群类型的基础上，能够更加细致地挖掘其主要特征。

此外，从其后的动态标签可以发现，“佳片推荐”长期的兴趣点为“爱情”和“喜剧”电影，但依然存在变化兴趣点。如 2018 年 1 月、2 月除“爱情”“喜剧”外较多关注“人性”方面的电影，2018 年 12 月则更多关注“青春”主题的电影。这是由于不同主题的电影热映，激起社群成员的短暂兴趣，故随之改变。

而对于“一个人看电影”长期兴趣为“爱情”，分析其社群名称可以猜测单身的社群成员依然对爱情有着非常强烈的向往。同时通过对 2018 年 1 月前后上映电影进行查询，《水形物语》《三块广告牌》《无问西东》等国内外经典影片都在 1 月份前后上映，这些电影都与“人性”有关，因此两个社群都在 2018 年 1 月、2 月逐渐开始讨论人性方面的电影问题。而 2018 年 12 月在中国上映的《狗十三》激起社会对青少年成长、青春主题的关注，引起全社会的热烈讨论，因此两个电影社群都在 12 月份生成“青春”标签。

因此，根据对同类型不同社群同一时间点所生成的动态标签进行对比，可以看出模型生成的两个社群的关注点和兴趣点是有所不同的。而针对不同类型的社群模型也能准确地识别，“买书 读书 一起来吧”模型识别其主要特征为“书籍”“买书”，并在生成的动态标签中将其社群成员对书籍的兴趣点进行表征。可以认为，模型能够对任意新增社群从其现有数据中挖掘其社群类型和动态社群兴趣，简化网络社交平台的管理和网络用户的使用。

5 结语

综上所述，本研究所提出的模型将社群话题表征的社群长期特征与社群活跃成员兴趣标签表征的社群短期兴趣进行结合，能够较好地揭示社群关注的特点。对社群标签的动态生成能够提高网络社群定义的及时性与准确性，方便用户清楚地了解不同社群特点，解决用户获取信息、选择社群困难等问题。但是，由于豆瓣用户多是使用概括性或反映整体感受、评价的标签[15]，因此在表征社群成员兴趣时有些标签的区分度不高，但依然能够依据现实情况及时地对社群成员兴趣予以表征。后续在数据更为合理的情况下，模型能够更准确及时地为社群生成表征其特点兴趣的标签。

参考文献：

[1] 邓胜利,胡吉明. Web 2.0 环境下网络社群理论研究综述[J]. 中国图书馆学报,2010,36(5): 90 – 95.
[2] RHEINGOLD H. The virtual community: finding commection in a computerized world[M]. Boston:Addison-wesley longman publishing co. , Inc. , 1993.
[3] SIEMENS G, BAKER R S J D. Learning analytics and educational data mining: towards communication and collaboration[C]// Proceedings of the 2nd international conference on learning analytics and knowledge. Vancouver, BC:ACM, 2012: 252 – 254.
[4] NAVARRO N D B, DA COSTA C A, BARBOSA J L V, et al. Spontaneous social network: toward dynamic virtual communities based on context-aware computing[J]. Expert systems with applications,2018,95:72 – 87.
[5] ZHOU T. Understanding online knowledge community user continuance: a social cognitive theory perspective[J]. Data technologies and applications,2018,52(3): 445 – 458.
[6] 刘佩,林如鹏. 网络问答社区"知乎"的知识分享与传播行为研究[J]. 图书情报知识,2015(6):109 – 119.
[7] 邓卫华,闫明星,易明. LPP 视角下网络社区用户口碑信息传播行为研究[J]. 情报资料工作,2017(1):82 – 87.
[8] LIAO C C, TO P L, HSU F C,. Exploring knowledge sharing in virtual communities[J]. Online information review,2013,37(6): 891 – 909.
[9] CHEN C, DU R, LI J, et al. The impacts of knowledge sharing-based value co-creation on user continuance in online communities [J]. Information discovery and delivery,2017,45(4):227 – 239.
[10] XIE H R, LI Q, MAO X D, et al. Mining latent user community for tag-based and content-based search in social media[J]. Computer journal, 2014, 57(9):1415 – 1430.
[11] 李文根. 基于社区问答系统的中文短文本标签生成研究[D]. 南京:南京大学,2017.
[12] 蒋武轩,熊回香,叶佳鑫,等. 网络社交平台中社群标签动态生成研究[J]. 数据分析与知识发现,2019,3(10):98 – 109.
[13] 李雷,朱玉婷,施化吉,等. 社会网络中基于 U_BTM 模型的主题挖掘[J]. 计算机应用研究,2017,34(1):132 – 135,146.
[14] 李敬,印鉴,刘少鹏,等. 基于话题标签的微博主题挖掘[J]. 计算机工程,2015,41(4):30 – 35.
[15] 林鑫,周知. 用户认知对标签使用行为的影响分析——基于电影社会化标注数据的实证分析[J]. 情报理论与实践,2015,38(10):85 – 88.

作者贡献说明：

蒋武轩:数据收集、模型构建、论文撰写与修改；
易明:论文指导；
熊回香:论文指导；
童兆莉:论文修改。

Research on the Generation of Community Tags in Network Social Platform

Jiang Wuxuan　Yi Ming　Xiong Huixiang　Tong Zhaoli
School of Information Management, Central China Normal University, Wuhan 430079

Abstract: [Purpose/significance] Community tags generated based on the mining of community topics and users' interests in network social platforms can improve the timeliness and accuracy of the definition of community, and solve the difficulties of user information acquisition and network community selection. [Method/process] Through in-depth analysis of the network community, it was determined that the community features can be represented according to the community topics and users' interests. Firstly, the BTM model of topic extraction was used to train the topic model of network social topics, and the pre-label of network social topics was obtained. Then, based on the different important node indexes of community members' interest tag network, the TOPSIS multi-index comprehensive evaluation method was used to mine the overall interest of members, so as to obtain the interest pre-label of members of the network community. After combining the two results, the community tag was generated and optimized. And this paper took "Douban Group" as an example for demonstration. [Result/conclusion] The community tag generation model based on community topics and members' interests can accurately mine the main interests and recent concerns. Tag generation of the community as a whole is conducive to the selection of interest groups of network users.

Keywords: community labels　tag generation　BTM　TOPSIS